Delving into the Convergence of Generalized Smooth Minimax Optimization

Wenhan Xian [1] **Ziyi Chen** [1] **Heng Huang** [1]

Abstract

Minimax optimization is fundamental and important for enormous machine learning applications such as generative adversarial network, adversarial training, and robust optimization. Recently, a variety of minimax algorithms with theoretical guarantees based on Lipschitz smoothness have been proposed. However, these algorithms could fail to converge in practice because the requisite Lipschitz smooth condition may not hold even in some classic minimax problems. We will present some counterexamples to reveal this divergence issue. Thus, to fill this gap, we are motivated to delve into the convergence analysis of minimax algorithms under a relaxed Lipschitz smoothness condition, *i.e.*, generalized smoothness. We prove that variants of basic minimax optimization algorithms GDA, SGDA, GDmax and SGDmax can still converge in generalized smooth problems, and hence their theoretical guarantees can be extended to a wider range of applications. We also conduct a numerical experiment to validate the performance of our proposed algorithms.

1. Introduction

The minimax problem is attracting growing attention due to its widespread practical applications in machine learning such as Generative Adversarial Net (GAN) (Goodfellow et al., 2014), adversarial training (Madry et al., 2017), robust optimization (Chen et al., 2017) and AUC maximization (Gao et al., 2013). In minimax optimization, variable x aims to minimize a pay-off loss function $f(x,y):\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ while variable y tries to maximize the loss, which can be formulated as

$$\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathcal{Y}} f(x,y), \tag{1}$$

where $\mathcal{Y}\subseteq\mathbb{R}^{d_2}$ is a convex domain. In this paper we consider the nonconvex strongly-concave problem where $f(x,y)$ in nonconvex in x and strongly-concave in y. In this case, the maximizer $y^*(x)=\arg\max_{y\in\mathcal{Y}} f(x,y)$ is unique and the primal objective function $\Phi(x)=f(x,y^*(x))$ can be well defined. The convergence criterion is to find a first-order stationary point of $\Phi(x)$ such that $\|\nabla\Phi(x)\|\le\epsilon$ for some tolerance ϵ. When considering stochastic problems, function $f(x,y)$ takes the form $f(x,y)=\mathbb{E}_{\xi\sim D}F(x,y;\xi)$ where $F(x,y;\xi)$ is the component loss function regarding sample ξ and D is the data distribution.

In recent years, minimax optimization problem is studied in a variety of research fields. Many deterministic and stochastic gradient-based methods with non-asymptotic convergence analysis for nonconvex strongly-concave minimax problems have been developed. Among these methods, some algorithms adopt the single-loop structure that updates x and y at the same frequency, such as Gradient Descent Ascent (GDA) and Stochastic Gradient Descent Ascent (SGDA) (Lin et al., 2020a). Some algorithms update x and y at different frequencies which involves a nested loop to search the optimal value of the maximizer y for the given x. Classic examples of double-loop minimax algorithms are GDmax and SGDmax (Jin et al., 2020). Some methods adopt more sophisticated structures to pursue better theoretical results (Lin et al., 2020b; Yang et al., 2020). Besides, some works also investigate the lower bound estimation of minimax problems (Li et al., 2021; Zhang et al., 2021) and some algorithms have been proven to be optimal or near-optimal (Lin et al., 2020b).

Although gradient-based minimax optimization algorithms have achieved huge success in theoretical region, most of the analysis frameworks are based on the requirement of Lipschitz smoothness. Some works conduct the convergence analysis without the Lipschitz smooth assumption for convex or weakly-convex problems (Rafique et al., 2022) and achieve competitive results, but the investigation for nonconvex generalized smooth minimax optimization is still limited. This drawback will restrict the applications of minimax optimization algorithms because in some cases the minimax structure breaks the Lipschitz smooth condition such as distributionally robust optimization (Yan et al., 2019; Levy et al., 2020; Jin et al., 2021), and in some machine learning tasks the objective function itself does not

[1]Department of Computer Science, University of Maryland, College Park, MD, United States. Correspondence to: Wenhan Xian <wxian1@umd.edu>, Ziyi Chen <zc286@umd.edu>, Heng Huang <heng@umd.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

satisfy the Lipschitz smoothness such as phase retrieval (Drenth, 2007; Miao et al., 1999). Counterexamples will be demonstrated in Section 2 to illustrate the divergence issue. Therefore, to fill this gap, we are motivated to investigate the convergence analysis of minimax algorithms under the relaxation of Lipschitz smooth assumption so that these algorithms can be theoretically guaranteed to work for a wider range of applications.

We summarize our contribution as follows.

- In this paper we study the convergence analysis of minimax optimization algorithms without the assumption of Lipschitz smoothness. We provide some counterexamples to reveal the divergence issue and propose the strategy to solve this problem.
- We prove that generalizations of classic minimax optimization algorithms (including single-loop algorithms GDA, SGDA and double-loop algorithms GDmax, SGDmax) can still converge under the generalized smooth condition and the gradient complexity matches the Lipschitz smooth counterparts. We conduct a numerical experiment of robust logistic regression task to validate the practical performance of our method.

2. Preliminary

2.1. Minimax Optimization Algorithms

In recent years, many algorithms were proposed to solve the optimization of minimax, and many of them were studied under the nonconvex-strongly-concave condition. GDmax and its stochastic variant SGDmax (Jin et al., 2020) are representatives of double-loop minimax algorithms. In each iteration they compute the estimation of the maximizer $y_{t+1} \approx y^*(x_t)$ via a nested loop and then update $x_{t+1} = x_t - \eta_x \nabla_x f(x_t, y_{t+1})$. GDmax can reach a first-order stationary point with $O(\kappa^2\epsilon^{-2}\log(1/\epsilon))$ iterations, where $\kappa = L/\mu$ is the condition number, L is the Lipschitz constant and μ is the strong concavity constant. SGDmax achieves the stochastic first-order oracle (SFO) complexity of $O(\kappa^3\epsilon^{-4}\log(1/\epsilon))$ to achieve a first-order stationary point. GDA and its stochastic variant SGDA (Lin et al., 2020a) are representatives of single-loop minimax algorithms. In each iteration, they compute the partial derivatives with respect to x and y, respectively. Then variables x and y are updated by $x_{t+1} = x_t - \eta_x \nabla_x f(x_t, y_t)$ and $y_{t+1} = y_t + \eta_y \nabla_y f(x_t, y_t)$. GDA reaches a first-order stationary point with $O(\kappa^2\epsilon^{-2})$ iterations, SGDA achieves the SFO complexity of $O(\kappa^3\epsilon^{-4})$ to achieve a first-order stationary point. These algorithms are fundamental optimizers to solve minimax optimization problem and hence we will conduct convergence analysis based on these algorithms. More recently, some algorithms have been proposed to to accelerate the convergence rate and reduce the gradient complexity of minimax optimization by variance reduction, such as SREDA ((Luo et al., 2020)) and Acc-MDA ((Huang et al., 2022)). Moreover, on deterministic setting some recently proposed algorithms ((Lin et al., 2020b)) have already matched the optimal lower bound ((Zhang et al., 2021)).

2.2. Counterexamples in Minimax Problems

In this section we will provide some counterexamples to illustrate the non-Lipschitz smoothness and divergence issue in minimax optimization. First we will introduce some basic definitions about Lipschitz smoothness.

Definition 2.1. A real-value function f is Lipschitz smooth if there exists a constant L such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{2}$$

Definition 2.2. A real-value function f is Lipschitz continuous if there exists a constant M such that

$$\|f(x) - f(y)\| \leq M\|x - y\| \tag{3}$$

Example 1. We will take distributionally robust optimization as our first example, which is a classic application of minimax optimization. Distributionally robust optimization aims to make the training result of the original optimization problem more robust by introducing a perturbation and solving a minimax problem. In (Yan et al., 2019), an example of this task is formulated as

$$\min_x \max_{y\in\Delta_n} f(x, y) = \sum_{i=1}^n y_i l_i(x) - V(y) \tag{4}$$

where n is the number of samples and $l_i(x)$ is the original loss function. Δ_n is the simplex in n-dimensional Euclidean space and $V(y)$ denotes a divergence measure between two distributions, which could be chosen as $\sum_{i=1}^n (y_i - \frac{1}{n})^2$. In this case, we can see Problem (4) is a nonconvex-strongly-concave minimax problem. We assume that original loss functions $l_i(x)$ are Lipschitz smooth but not Lipschitz continuous. Then we have

$$\|\nabla_y f(x, y) - \nabla_y f(x', y)\|^2 = \sum_{i=1}^n (l_i(x) - l_i(x'))^2 \tag{5}$$

If function f is Lipschitz smooth, we should have

$$\|\nabla_y f(x, y) - \nabla_y f(x', y)\|^2 \leq L^2\|x - x'\|^2 \tag{6}$$

which implies each $l_i(x)$ is Lipschitz continuous and conflicts with our assumption. Hence the objective function f is not Lipschitz smooth even the original loss functions l_i are Lipschitz smooth, which shows that the minimax structure can probably break the condition of Lipschitz smoothness.

The convergence analysis of most current existing minimax algorithms is based on the Lipschitz smoothness assumption.

However, this condition is not satisfied in many classic examples such as robust optimization. This result motivates us to study the convergence of minimax algorithms without the requirement of Lipschitz smoothness.

Example 2. Next we will provide a simple example to reveal the divergence issue when Lipschitz smoothness is not satisfied. We define a minimax problem

$$\min_x \max_y f(x,y) = yx^2 - 0.5y^2 \tag{7}$$

where x and y are scalars. It is easy to check $y^*(x) = x^2$ and $\Phi(x) = 0.5x^4$. Thus, we have $\nabla\Phi(x) = 2x^3$. For any fixed stepsize $\eta > 0$, if we choose initial value $x_0 \geq \frac{2}{\sqrt{\eta}}$ and apply a gradient descent algorithm, then we can prove $|\eta\nabla\Phi(x_t)| \geq |x_t|$ and $|x_{t+1}| \geq 2|x_t|$ for all $t \geq 0$. It implies that $|x_t| \geq 2^t|x_0|$ and the algorithm will diverge. In this paper, we will some generalized minimax algorithms to tackle the divergence issue.

2.3. Generalized Smoothness

Previous works studying nonconvex nonsmooth minimax optimization can be categorized into following branches. Some minimax algorithms adopt the zeroth-order strategy (Liu et al., 2019; Wang et al., 2020; Huang et al., 2022) to address issue where the objective function is not differentiable or the gradient cannot be accessed. However, if the objective function is still differentiable, just not Lipschitz smooth, gradient-based methods are more efficient and effective than gradient-free methods. Some other works focus on nonconvex nonsmooth minimax problems with certain special structures. As an example, (Huang et al., 2021) considers the problem that is a nonconvex Lipschitz smooth loss function adding a convex nonsmooth regularization, which can be solved by proximal gradient. (Li et al., 2022) considers a nonsmooth composite minimax problem where $f(\cdot, y)$ is the composition of a Lipschitz smooth function and Lipschitz continuous function. In this paper, we do not assume any specific structures for the objective function.

In a concurrent work (Hao et al., 2024), the convergence analysis of a bilevel optimization algorithm under the condition of unbounded smoothness is provided, which is also applicable to minimax optimization. In (Hao et al., 2024), the lower level function that is used to calculate $y^*(x)$ is assumed to be Lipschitz smooth and the upper level function is assumed to be (L_0, L_1)-smooth (Zhang et al., 2019), which is defined as follows:

Definition 2.3. A real-value function f is (L_0, L_1)-smooth if there exist constants L_0 and L_1 such that

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1\|\nabla f(x)\|)\|x - y\| \tag{8}$$

We can see Lipschitz smoothness is a special case of (L_0, L_1)-smoothness where $L_1 = 0$. Recently, a variety of works are proposed to study and generalize the requirement of Lipschitz smoothness (Chen et al., 2023; Li et al., 2023). In (Li et al., 2023), the definition of l-smoothness is proposed as follows:

Definition 2.4. A real-value function f is l-smooth if there exists a non-decreasing continuous function $l(\cdot)$ such that

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq l(\|\nabla f(x)\| + G) \cdot \|x_1 - x_2\| \tag{9}$$

for any x_1 and x_2 in $\mathcal{B}(x, \frac{G}{l(\|\nabla f(x)\|+G)})$ for any $G > 0$.

In (Li et al., 2023), it is proven that Definition 9 is equivalent to $\|\nabla^2 f(x)\| \leq l(\|\nabla f(x)\|)$ almost everywhere. For nonconvex optimization problems, function l is required to be sub-quadratic but (L_0, L_1)-smoothness still can be regarded as a special case of l-smoothness where $l(u) = L_0 + L_1 u$. A common example of sub-quadratic function is $l(u) = L_0 + L_\rho u^\rho$ where $0 < \rho < 2$, which contain the case of $\rho = 1$. In this paper, we extend the concept of l-smoothness to minimax optimization and propose the definition of l_x-l_y-smoothness in Definition 2.5. Therefore, the smoothness condition used in this paper is more general than the assumption used in (Hao et al., 2024).

Definition 2.5. A real-value function $f(x, y) : \mathbb{R}^{d_1} \times \mathcal{Y} \to \mathbb{R}$ is called l_x-l_y-smooth for non-decreasing continuous functions l_x and l_y if we have

$$\|\nabla_x f(z_1) - \nabla_x f(z_2)\| \leq l_x(\|\nabla_x f(z_0)\| + G_1) \cdot \|z_1 - z_2\|$$
$$\|\nabla_y f(z_1) - \nabla_y f(z_2)\| \leq l_y(\|\nabla_y f(z_0)\| + G_2) \cdot \|z_1 - z_2\|$$

for any z_1 and z_2 in $\mathcal{B}(z_0, r(z_0))$ and any $z_0 = [x_0; y_0]$, where $r(z_0) = \frac{G_1}{l_x(\|\nabla_x f(z_0)\|+G_1)} + \frac{G_2}{l_y(\|\nabla_y f(z_0)\|+G_2)}$ for any given $G_1 > 0$ and $G_2 > 0$.

Moreover, it can be proven that the two counterexamples belong to the category of our l_x-l_y-smoothness.

Algorithm 1 Generalized GDA or SGDA

Input: initial value x_0 and y_0
Parameter: learning rate η and η_y, maximum iteration T.
for $t = 0, 1, \ldots, T-1$ **do**
  Compute $v_t = \nabla_x f(x_t, y_t)$ (deterministic) or $v_t = \nabla_x F(x_t, y_t; \xi_t)$ (stochastic).
  Compute $u_t = \nabla_y f(x_t, y_t)$ (deterministic) or $u_t = \nabla_y F(x_t, y_t; \xi_t)$ (stochastic).
  Compute suitable stepsize parameter S_t.
  Update $x_{t+1} = x_t - (\eta/S_t)v_t$.
  Update $y_{t+1} = \Pi_{\mathcal{Y}}(y_t + \eta_y u_t)$.
end for

3. Algorithms

As revealed in our counterexample, vanilla gradient based algorithm fails to converge in minimax optimization when

Algorithm 2 Generalized GDmax or SGDmax

Input: initial value x_0 and y_0
Parameter: learning rate η and η_y, nested loop size K, maximum iteration T.

for $t = 0, 1, \ldots, T-1$ **do**
  Compute $v_t = \nabla_x f(x_t, y_t)$ (deterministic) or $v_t = \nabla_x F(x_t, y_t; \xi_t)$ (stochastic).
  Compute suitable stepsize parameter S_t.
  Update $x_{t+1} = x_t - (\eta/S_t)v_t$.
  Let $y_{t,0} = y_t$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Compute $u_{t,k} = \nabla_y f(x_{t+1}, y_{t,k})$ (deterministic) or $u_{t,k} = \nabla_y F(x_{t+1}, y_{t,k}; \xi_{t,k})$ (stochastic).
    Update $y_{t,k+1} = \Pi_{\mathcal{Y}}(y_{t,k} + \eta_y u_{t,k})$.
  end for
  Update $y_{t+1} = y_{t,K}$.
end for

the Lipschitz smooth assumption does not hold. The reason for divergence is credit to the large gradient. Therefore, we will generalize these algorithms to tackle this issue by adopting an suitable stepsize strategy to control the moving distance in each iteration. We will apply this strategy to standard minimax optimizers GDA, SGDA, GDmax and SGDmax. The description of single-loop algorithm Generalized GDA (or SGDA) is shown in Algorithm 1. The description of double-loop algorithm Generalized GDmax (or SGDmax) is shown in Algorithm 2.

Let x_0 and y_0 be the initial values in Algorithm 1 and Algorithm 2. In our convergence analysis, we need to run an additional initialization process to obtain an approximation of the maximizer $y_0 \approx y^*(x_0)$ for the given initial value x_0 before the algorithms start. The specific conditions that y_0 needs to satisfy will also be discussed in the convergence analysis This subproblem can be converted to a strongly-convex generalized Lipschitz smooth minimization problem and solved by optimizers such as GD, SGD or SPIDER (Chen et al., 2023; Li et al., 2023; Fang et al., 2018).

In Algorithm 1, we adopt a suitable stepsize based on the norm of gradient to single-loop minimax algorithms GDA and SGDA. In each iteration, we compute the gradients $\nabla_x f(x_t, y_t)$, $\nabla_y f(x_t, y_t)$ or the corresponding stochastic gradients with respect to x and y, respectively. Then we update x_t and y_t by gradient descent ascent. When we update x_t, we adopt the suitable stepsize strategy to control the moving distance. We have multiple options to compute the suitable stepsize parameter S_t. It could be:

(1) $S_t = \|v_t\|$. (2) $S_t = \max\{\epsilon, \frac{1}{t+1}\sum_{\tau=0}^{t}\|v_t\|\}$.

(3) $S_t \equiv S$. (4) $S_t = \max\{\epsilon, (1-\beta)\|v_t\| + \beta S_{t-1}\}$.

When we choose option (1), the suitable stepsize strategy is turned out to be the gradient normalization method. When we choose option (2), we calculate the average of historical gradient norm. When we choose option (3), the suitable stepsize will be a constant. Notice that it is different from the conventional constant stepsize because S probably has dependence on the initial value and it is calculated after the algorithm starts. When we choose option (4), we calculate the exponential average of historical gradient norm. When we update y_t, we adopt a constant stepsize such that $y_{t+1} = \Pi_{\mathcal{Y}}(y_t + \eta_y u_t)$. with a projection onto $\mathcal{Y}$.

In Algorithm 2, we apply the suitable stepsize strategy to double-loop minimax algorithms GDmax and SGDmax. In each iteration, we first compute the gradient $\nabla_x f(x_t, y_t)$ with respect to x (or the corresponding stochastic gradient). We update $x_{t+1} = x_t - (\eta/S_t)v_t$ by an suitable stepsize η/S_t, where the options to compute S_t are the same as Algorithm 1. Then we run a nested loop to search an estimation of the maximizer $y_{t+1} \approx y^*(x_{t+1})$. Specifically, we apply an iterative gradient ascent algorithm $y_{t,k+1} = \Pi_{\mathcal{Y}}(y_{t,k} + \eta_y u_{t,k})$ where $u_{t,k}$ is the deterministic or stochastic gradient estimator to solve the maximization subproblem $\max_y f(x_{t+1}, y)$.

4. Convergence Analysis

4.1. Main Theorems

In this section, we will show the main theorems of our convergence analysis. The theoretical results indicate that our generalized GDA, SGDA, GDmax or SGDmax algorithms can converge under the generalized Lipschitz smooth condition and the gradient complexities to reach first-order stationary point are the same as Lipschitz smooth counterparts. First we will introduce the following assumptions.

Assumption 4.1. The primal function Φ is lower bounded, *i.e.*, $\inf_x \Phi(x) = \Phi^* > -\infty$.

Assumption 4.2. The loss function $f(x, y)$ is μ-strongly-concave w.r.t. y, *i.e.*, there exists a constant $\mu > 0$ such that for any x, y and y', we have

$$f(x, y) \le f(x, y') + \langle \nabla_y f(x, y'), y - y' \rangle - \frac{\mu}{2}\|y - y'\|^2$$

Assumption 4.3. The loss function $f(x, y)$ is l_x-l_y-smooth and function l_x is sub-quadratic.

These assumptions are basic prerequisites for the convergence analysis of nonconvex strongly-concave minimax optimization. In nonconvex minimization problems (Li et al., 2023), the function l is also required to be sub-quadratic.

We conduct our convergence analysis based on two cases. The first case is $\mathcal{Y} = \mathbb{R}^{d_2}$, which results in an unconstrained optimization with respect to y. The second case is that $\mathcal{Y}$ is bounded, which implies f is Lipschitz smooth with respect

to y, *i.e.*, there exists a constant L_y such that $l_y(\cdot) \equiv L_y$. We need these requirements because otherwise the value of $l_y(\|\nabla_y f(x, y^*(x))\|)$ is hard to estimate, which can lead to poor smoothness even approaching the maximizer y^*.

We provide the following essential definitions of notations that are frequently used in our analysis.

$$G_x = \max\{u > 0 | u^2 \leq 8\kappa l_x(2u) \cdot (\Phi(x_0) - \Phi^*)\}$$
$$G_y = \nabla_y f(x_0, y_0),\ y_t^* = y^*(x_t) \tag{10}$$

4.1.1. Analysis Results of GDA

Similar to Lipschitz smooth minimax problems, we can define the condition number as $\kappa = l_y(4G_y)/\mu$. With Assumption 4.1 to 4.3, we can obtain the following Theorem for the generalized GDA algorithm.

Theorem 4.4. *Assume Assumption 4.1, 4.2 and 4.3 are satisfied. Let parameters $\frac{\eta}{S_t} \leq \frac{C_0}{16\kappa^2 l_x(2G_x)}$ for all t, $\eta_y = \frac{1}{l_y(2G_y)}$ and initial value $\|y_0 - y_0^*\| \leq \frac{C_0 G_x}{l_x(2G_x)}$ where constant $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(0)G_x}\}$. Then for the generalized GDA algorithm, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{5(\Phi(x_0) - \Phi^*)}{\eta T} \tag{11}$$

When S_t is constant (option (3)), we can achieve the following Corollary 4.5 for generalized GDA, which indicates that under the condition of generalized Lipschitz smoothness our generalized GDA algorithm can achieve the same gradient complexity to find first-order stationary point as GDA does with Lipschitz smoothness.

Corollary 4.5. *When S_t is computed by option (3), let $\eta = O(\frac{1}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $T = O(\kappa^2\epsilon^{-2})$ and other conditions are the same as Theorem 4.4. Then the generalized GDA algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2})$ gradient oracles.*

When we choose other options to compute S_t, we can obtain the following theoretical results.

Corollary 4.6. *When S_t is computed by option (1) or (4), let $\eta = O(\frac{\epsilon}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $T = O(\kappa^2\epsilon^{-2})$ and other conditions are the same as Theorem 4.4. Then the generalized GDA algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2})$ gradient oracles.*

Corollary 4.7. *When S_t is computed by option (2), let $\eta = O(\frac{\epsilon}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $T = O(\kappa^2\epsilon^{-2}\log(\frac{1}{\epsilon}))$ and other conditions are the same as Theorem 4.4. Then the generalized GDA algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2}\log(\frac{1}{\epsilon}))$ gradient oracles.*

We can see the gradient oracle complexity to achieve first-order stationary points is the same as GDA when the suitable stepsize parameter S_t is computed by gradient norm or exponential moving average of historical gradient norm. When S_t is computed by averaged historical gradient norm, there will be an additional logarithm term. However, our analysis is conducted under the condition of generalized Lipschitz smoothness, while the original analysis of GDA is based on Lipschitz smoothness.

4.1.2. Analysis Results of GDmax

For double-loop deterministic algorithm Generalized GDmax, we have the following Theorem 4.8.

Theorem 4.8. *Assume Assumption 4.1, 4.2 and 4.3 are satisfied. Let parameters $\frac{\eta}{S_t} \leq \frac{C_0}{16\kappa l_x(2G_x)}$ for all t, $\eta_y = \frac{1}{l_y(4G_y)}$, $K \geq \kappa\log(\frac{1}{\theta})$ and initial value $\|y_0 - y_0^*\| \leq \frac{C_0 G_x}{l_x(2G_x)}$ where constant $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$ and $\theta = \min\{\frac{1}{4}, \frac{1}{l_x^2(2G_x)}\}$. Then for the generalized GDmax algorithm, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{5(\Phi(x_0) - \Phi^*)}{\eta T} \tag{12}$$

Similar to generalized GDA, we can prove under the condition of generalized Lipschitz smoothness, GDmax algorithm can achieve the same gradient complexity to find first-order stationary point as GDmax does with Lipschitz smoothness.

Corollary 4.9. *When S_t is computed by option (3), let $\eta = O(\frac{1}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T = O(\kappa\epsilon^{-2})$ and other conditions are the same as Theorem 4.8. Then the generalized GDmax algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2})$ gradient oracles.*

Corollary 4.10. *When S_t is computed by option (1) or (4), let $\eta = O(\frac{\epsilon}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T = O(\kappa\epsilon^{-2})$ and other conditions are the same as Theorem 4.8. Then the generalized GDmax algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2})$ gradient oracles.*

Corollary 4.11. *When S_t is computed by option (2), let $\eta = O(\frac{\epsilon}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T = O(\kappa\epsilon^{-2}\log(\frac{1}{\epsilon}))$ and other conditions are the same as Theorem 4.8. Then Generalized GDmax algorithm can find an ϵ-first-order stationary point with $O(\kappa^2\epsilon^{-2}\log(\frac{1}{\epsilon}))$ gradient oracles.*

4.1.3. Analysis Results of SGDA

For generalized stochastic algorithms SGDA and SGDmax, we assume the stochastic gradient oracle is unbiased, *i.e.*, $\mathbb{E}_\xi \nabla F(x, y; \xi) = \nabla f(x, y)$. We also need the following bounded variance assumption, which is a common assumption in the convergence analysis of stochastic gradient-based optimization algorithms.

Assumption 4.12. The stochastic gradient oracle satisfies $\mathbb{E}_\xi\|\nabla F(x, y; \xi) - \nabla f(x, y)\|^2 \leq \sigma^2$ for some constant σ.

	Generalized GDA / SGDA	Generalized GDmax / SGDmax
(1)	$\eta = O(\frac{\epsilon}{\kappa^2}), T = O(\kappa^2\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$	$\eta = O(\frac{\epsilon}{\kappa}), T = O(\kappa\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$
(2)	$\eta = O(\frac{\epsilon}{\kappa^2}), T = \tilde{O}(\kappa^2\epsilon^{-2}), SFO = \tilde{O}(\kappa^3\epsilon^{-4})$	$\eta = O(\frac{\epsilon}{\kappa}), T = \tilde{O}(\kappa\epsilon^{-2}), SFO = \tilde{O}(\kappa^3\epsilon^{-4})$
(3)	$\eta = O(\frac{1}{\kappa^2}), T = O(\kappa^2\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$	$\eta = O(\frac{1}{\kappa}), T = O(\kappa\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$
(4)	$\eta = O(\frac{\epsilon}{\kappa^2}), T = O(\kappa^2\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$	$\eta = O(\frac{\epsilon}{\kappa}), T = O(\kappa\epsilon^{-2}), SFO = O(\kappa^3\epsilon^{-4})$

Table 1. A summary of the stepsize η, total iterations T with respect to Generalized GDA, SGDA, GDmax and SGDmax algorithms and different choices of S_t. Column one refers to different options to compute S_t, which is defined in Section 3. SFO refers to the stochastic first-order oracle for stochastic algorithms SGDA and SGDmax. Notation $\tilde{O}(\cdot)$ hides the logarithm term.

In stochastic algorithms, let b_x and b_y denote the batchsize of stochastic gradient with respect to x and y, respectively. Due to the noise of stochastic gradient, there is no guarantee for the upper bound of gradient or function value. Thus, we cannot apply mathematical induction to estimate the upper bound along the trajectory, as what we do in the deterministic case (see the sketch of proof in next subsection). However, we can still prove that generalized SGDA and SGDmax will converge with a high probability. In the stochastic case, we need to re-define the constant

$$G_x = \max\{u > 0 | u^2 \leq 32\kappa l_x(2u)\cdot(\Phi(x_0) - \Phi^* + \sigma^2)/\delta\}$$

For Generalized SGDA, we have the following Theorem.

Theorem 4.13. *Assume Assumption 4.1, 4.2, 4.3 and 4.12 are satisfied. Let parameters* $\frac{\eta}{S_t} \leq \frac{\delta C_0}{48\kappa^2 l_x(2G_x)}$ *for all* t, $\eta_y = \frac{1}{l_y(2G_y)}$, $T = \frac{\kappa^2}{\delta^2\epsilon^2}$, $b_x \geq \frac{\sigma^2}{G_x^2\epsilon^2}$, $b_y \geq \max\{\frac{192\kappa\sigma^2 l_x^2(2G_x)}{\delta G_x^2 l_y^2(2G_y)}, \frac{\kappa l_x(2G_x)}{\delta^2 l_y^2(2G_y)\epsilon^2}\}$ *and initial value* $\|y_0 - y_0^*\| \leq \frac{\delta C_0 G_x}{8 l_x(2G_x)}$ *where constant* $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$. *Then for the generalized SGDA algorithm, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{10(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta\eta T} \tag{13}$$

with probability at least $1 - \delta$.

When S_t is constant ($S_t \equiv S$), we can obtain the following Corollary for Generalized SGDA, which results in the same stochastic first-order oracle complexity under the condition of relaxed Lipschitz smoothness as SGDA does with the requirement of Lipschitz smoothness.

Corollary 4.14. *When S_t is computed by option (3), let $\eta = O(\frac{1}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $b_x = O(\epsilon^{-2})$, $b_y = O(\kappa\epsilon^{-2})$, $T = O(\kappa^2\epsilon^{-2})$ and other conditions are the same as Theorem 4.13. Then the generalized SGDA algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4})$.*

When S_t is computed by option (1) or (4), we can reach the following conclusion which also achieves the same SFO complexity as SGDA does in the Lipschitz smooth case.

Corollary 4.15. *When S_t is computed by option (1) or (4), let $\eta = O(\frac{\epsilon}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $b_x = O(\epsilon^{-2})$, $b_y = O(\kappa\epsilon^{-2})$, $T = O(\kappa^2\epsilon^{-2})$ and other conditions are the same as Theorem 4.13. Then the generalized SGDA algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4})$.*

When S_t is computed by option (4), we can obtain the following theoretical result, which causes an additional logarithm term in the SFO complexity.

Corollary 4.16. *When S_t is computed by option (2), let $\eta = O(\frac{\epsilon}{\kappa^2 l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(2G_y)})$, $b_x = O(\epsilon^{-2})$, $b_y = O(\kappa\epsilon^{-2})$, $T = O(\kappa^2\epsilon^{-2}\log(\frac{1}{\epsilon}))$ and other conditions are the same as Theorem 4.13. Then the generalized SGDA algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4}\log(\frac{1}{\epsilon}))$.*

4.1.4. Analysis Results of SGDmax

For the stochastic double-loop algorithm Generalized SGDmax, we have the following conclusions.

Theorem 4.17. *Assume Assumption 4.1, 4.2, 4.3 and 4.12 are satisfied. Let parameters* $\frac{\eta}{S_t} \leq \frac{\delta C_0}{48\kappa l_x(2G_x)}$ *for all* t, $\eta_y = \frac{1}{l_y(4G_y)}$, $K \geq \kappa\log(\frac{1}{\theta})$, $T = \frac{\kappa}{\delta^2\epsilon^2}$, $b_x \geq \frac{\sigma^2}{G_x^2\epsilon^2}$, $b_y \geq \max\{\frac{24\kappa\sigma^2 l_x^2(2G_x)}{\delta G_x^2 l_y^2(4G_y)}, \frac{\kappa l_x(2G_x)}{\delta^2 l_y^2(4G_y)\epsilon^2}\}$ *and initial value* $\|y_0 - y_0^*\| \leq \frac{\delta C_0 G_x}{8 l_x(2G_x)}$ *where constant* $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$ *and* $\theta = \min\{\frac{1}{4}, \frac{1}{l_x^2(2G_x)}\}$. *Then for the generalized SGDmax algorithm, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{10(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta\eta T} \tag{14}$$

with probability at least $1 - \delta$.

Corollary 4.18. *When S_t is computed by option (3), let $\eta = O(\frac{1}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T = O(\kappa\epsilon^{-2})$ and other conditions are the same as Theorem 4.17. Then the generalized SGDmax algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4})$.*

Corollary 4.19. *When S_t is computed by option (1) or (4), let $\eta = O(\frac{\epsilon}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T =$*

Name	a9a	covtype	diabetes	german	gisette	ijcnn1	mushrooms	phishing	w8a
Samples	32561	581012	768	1000	6000	141691	8124	11055	49749
Features	123	54	8	24	5000	22	112	68	300

Table 2. Descriptions of the LIBSVM binary classification datasets used in our experiment

$O(\kappa\epsilon^{-2})$ and other conditions are the same as Theorem 4.17. Then the generalized SGDmax algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4})$.

Corollary 4.20. *When S_t is computed by option (2), let $\eta = O(\frac{\epsilon}{\kappa l_x(2G_x)})$, $\eta_y = O(\frac{1}{l_y(4G_y)})$, $K = O(\kappa)$, $T = O(\kappa\epsilon^{-2}\log(\frac{1}{\epsilon}))$ and other conditions are the same as Theorem 4.17. Then the generalized SGDmax algorithm can find an ϵ-first-order stationary point with SFO of $O(\kappa^3\epsilon^{-4}\log(\frac{1}{\epsilon}))$.*

These theoretical results indicate that under the generalized Lipschitz smooth condition, our generalized SGDmax method can still converge and achieve the same SFO complexity as SGDmax does in the Lipschitz smooth case.

4.2. Sketch of Proof

In this subsection, we will provide the outline of our proof to illustrate the insight of our analysis. The completed proof is left to the Appendix. Due to the space limit, we will only demonstrate the sketch of proof for the generalized GDA and SGDA algorithms. First, we can prove the smoothness for functions $y^*(x)$ and $\Phi(x)$ (described in Lemma A.1 and Lemma A.2) such that $\|y^*(x) - y^*(x')\| \leq \kappa\|x - x'\|$ and

$$\|\nabla\Phi(x) - \nabla\Phi(x')\| \leq 2\kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x - x'\|$$

if $\|x' - x\| \leq \frac{G}{l_x(\|\nabla_x f(x,y^*(x))\|+G)}$ for some $G \geq 0$. Then we can obtain Lemma A.3, which indicates that

$$\|\nabla\Phi(x)\|^2 \leq 4\kappa l_x(2\|\nabla\Phi(x)\|) \cdot (\Phi(x) - \Phi^*) \qquad (15)$$

for $\forall x$. When function l_x is sub-quadratic, Eq. (15) provides an upper bound for $\|\nabla\Phi(x)\|$ that has dependence on the function value gap $(\Phi(x) - \Phi^*)$.

Next, we want to prove that $\|\nabla\Phi(x_t)\| \leq G_x$ for all $t \geq 0$ in Generalized GDA, which means the gradient is bounded along the trajectory x_t. With this conclusion, the values of $l_x(\cdot)$ that occur along the trajectory in the analysis can be bounded by $l_x(2G_x)$, and hence the rest part of the proof will be simplified and relatively easy. In minimization optimization, this conclusion can be directly achieved by mathematical induction. However, in mimimax optimization the exact value of $\nabla\Phi(x)$ is not available. It is estimated by $\nabla_x f(x_t, y_t)$, which yields an error term caused by $\|y_t - y_t^*\|$. The original proof framework of minimization problem does not work in this case due to the existence of the error term. Besides, the error term will lead to an additional term that also has dependence on G_x when bounding the function value gap $(\Phi(x) - \Phi^*)$. To solve this issue, we need to apply mathematical induction to $\nabla\Phi(x_t)$, $\nabla_x f(x_t, y_t)$, $\nabla_y f(x_t, y_t)$ and $\|y_t - y_t^*\|$ simultaneously to estimate the bound for these terms. This is one of the most challenging technical difficulties in our analysis. We can prove

$$\Phi(x_t) - \Phi^* \leq \Phi(x_0) - \Phi^* + \frac{G_x^2}{8\kappa l_x(2G_x)} \qquad (16)$$

which will eventually finalize the mathematical induction.

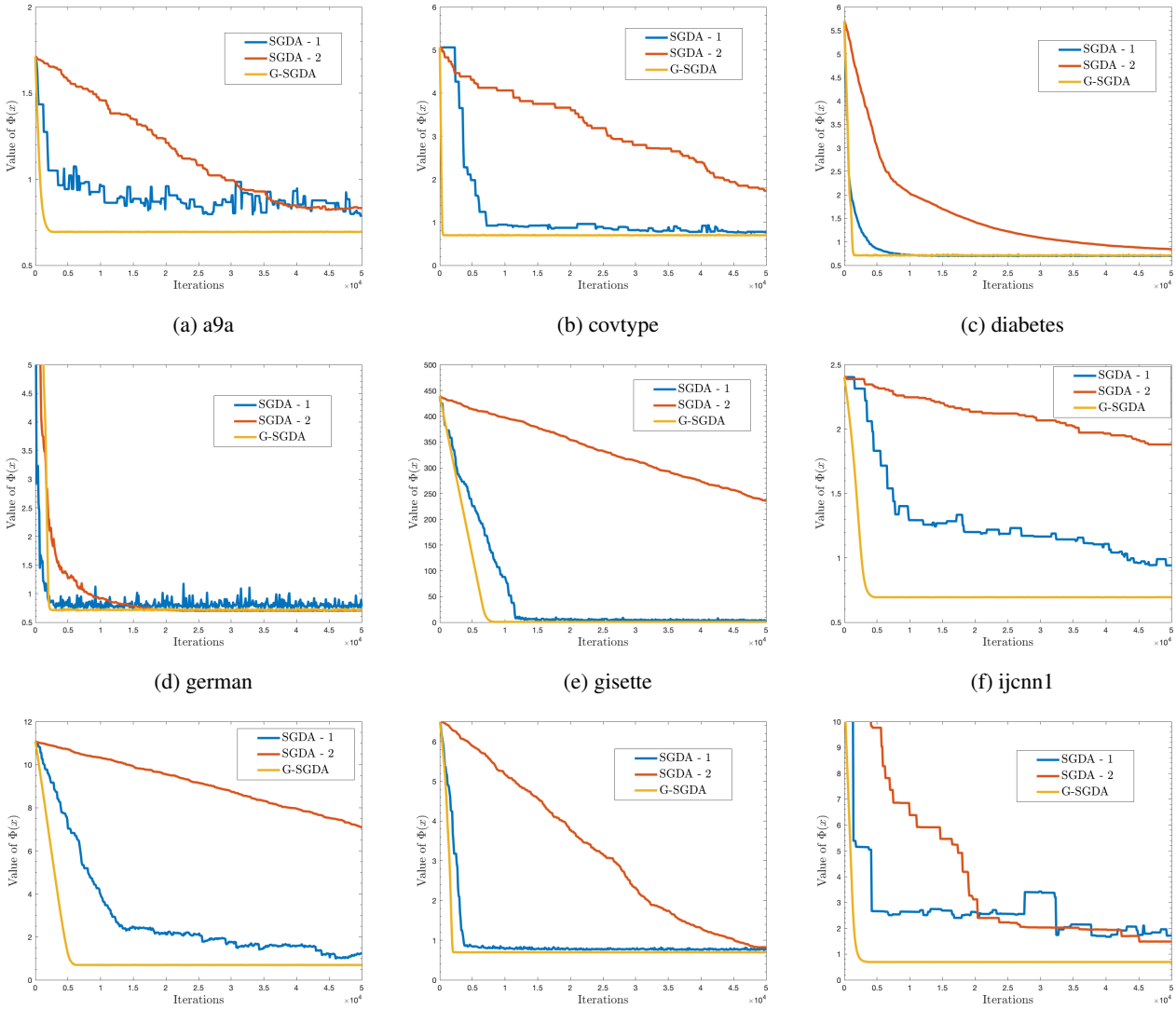
In the stochastic case, the framework of mathematical induction in GDA does not work because the neither the gradient norm nor the function value can be bounded when gradient noise exists. However, under these conditions we can still prove the convergence of our Generalized SGDA with a probability at least $1 - \delta$. For Generalized SGDA, we define

$$G_x = \max\{u > 0 | u^2 \leq 32\kappa l_x(2u) \cdot (\Phi(x_0) - \Phi^* + \sigma^2)/\delta\}$$
$$T_0 = \min\{t | \Phi(x_t) - \Phi^* > F \text{ or } \|y_t^* - y_t\| > Y\} \wedge T$$

where $F = 8(\Phi(x_0) - \Phi^* + \sigma^2)/\delta$, $Y = \frac{C_0 G_x}{l_x(2G_x)}$ and $\wedge$ denotes the minimum operation. We want to prove the probability of $T_0 < T$ is small. Notice that in minimization optimization we do not need to consider the upper bound of $\|y_t^* - y_t\|$, which is exclusive in minimax optimization. Based on the proof of the deterministic case we can prove when $t < T_0$ all induction assumptions in the analysis of GDA are satisfied. Hence we can obtain the estimations of expectations $\mathbb{E}\Phi(x_t) - \Phi^*$ and $\mathbb{E}\|y_t^* - y_t\|$ at iteration $t = T_0$. By Markov's inequality and union bound, we can prove the probability of event $T_0 < T$ is smaller than $\frac{\delta}{2}$. Furthermore, by union bound and the estimation of $\mathbb{E}\Phi(x_t)$ we can achieve the result in Theorem 4.13.

4.3. Discussion

In this subsection, we will discuss the dependence of constants used in our convergence analysis. Since we run an additional initialization process to ensure $\|y_0 - y_0^*\| \leq C$ for some threshold C, we can obtain $G_y \leq \frac{1}{4}$ if there is no constraint with respect to y, *i.e.*, $\mathcal{Y} = \mathbb{R}^{d_2}$. Thus, we have $\kappa \leq \frac{l_y(1)}{\mu}$. If f is Lipschitz smooth with respect to y, we also have $\kappa \leq \frac{l_y(1)}{\mu}$. Hence the condition number κ is a constant only depending on the function $l_y(\cdot)$. Insert $\kappa \leq \frac{l_y(1)}{\mu}$ into the definition of G_x, we can see G_x is a constant only



(a) a9a (b) covtype (c) diabetes

(d) german (e) gisette (f) ijcnn1

(g) mushrooms (h) phishing (i) w8a

Figure 1. Experimental results of the loss function value of $\Phi(x)$ with respect to the number of iterations in the robust logistic regression task on dataset a9a, covtype, diabetes, german, gisette, ijcnn1, mushrooms, phishing, and w8a. SGDA-1 and SGDA-2 are the results of SGDA with two largest learning rates that make it converge. G-SGDA is the result of our Generalized SGDA Algorithm.

depending on functions $l_x(\cdot), l_y(\cdot), \Phi(\cdot)$ and the initial value x_0. Besides, the initialization process can be regarded as a strongly-convex minimization subproblem, which aims to find an initial value satisfying $\|y_0 - y_0^*\|$ smaller than a constant tolerance. The complexity of this subproblem is proved to be within $O(\frac{l_y(2\|\nabla_y f(x_0,\tilde{y}_0)\|)}{\mu})$ where $\tilde{y}_0$ is the raw input of variable y. Therefore, the complexity of the initialization process is dominated by the complexity to solve the entire minimax problem and thus can be neglected.

Next, we will discuss the relation between parameters η and S_t. η can be regarded as a fixed stepsize parameter which is passed to the algorithm before it starts. S_t is the scale of suitable stepsize in iteration t which is computed during the runtime of the algorithm. The ratio of $\frac{\eta}{S_t}$ should be bounded by a certain threshold according to our analysis. When S_t is chosen as a constant, parameter η can also be a constant that has no dependence on ϵ. When S_t is the gradient norm, averaged historical gradient norm or exponential moving averaged historical gradient norm, parameter η should be as small as $O(\epsilon)$ with respect to ϵ because S_t will become as small as $O(\epsilon)$ gradually. A summary of the stepsize parameter η, total iterations T and stochastic first-order oracle complexity with respect to different algorithms and choices of S_t is shown in Table 1.

5. Experiments

In this section, we will conduct an experiment of the robust logistic regression task to validate the performance of our generalized minimax optimization methods with the suitable stepsize strategy. Recall the examples we have mentioned in Section 2, the problem can be formulated as

$$\min_{x\in\mathbb{R}^d}\max_{y\in\Delta_n} f(x,y)=\sum_{i=1}^{n} y_i l_i(x) - V(y) + g(x) \quad (17)$$

where $l_i(x)$ is the logistic loss function defined by $l_i(x) = \log(1+\exp(-b_i a_i^T x))$. $V(y)$ is a divergence measure defined by $V(y)=\frac{1}{2}\lambda_1\|ny-\mathbf{1}\|^2$. Notation Δ_n represents the simplex in $\mathbb{R}^n$, that is

$$\Delta_n=\{y\in\mathbb{R}^n|0\le y_i\le 1,\sum_{i=1}^{n} y_i=1\} \quad (18)$$

Function $g(x)$ is the regularization term that takes the form $g(x)=\lambda_2\sum_{i=1}^{d}\frac{\alpha x_i^2}{1+\alpha x_i^2}$. Following the experimental settings in (Yan et al., 2019), we set $\lambda_1=\frac{1}{n^2}$, $\lambda_2=0.001$ and $\alpha=10$ in our experiment.

We run the experiment and verify our method on 9 real-world datasets a9a, covtype, diabetes, german, gisette, ijcnn1, mushrooms, phishing, and w8a, which can be downloaded from the LIBSVM repository at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets. These datasets are frequently used in binary classification tasks. The description of these datasets is listed in Table 2.

We compare our generalized SGDA algorithm with suitable stepsize to the conventional constant stepsize SGDA. We choose option (1) to compute the suitable stepsize parameter S_t, which adopts the gradient normalization. The mini-batch size is set to 50. For each algorithm, we choose the best learning rates η and η_y from $\{0.1, 0.01, 0.001, 0.0001, 1e-5, 1e-6\}$ by grid search. We report the results of two largest learning rates that can make SGDA converge. We compare the value of $\Phi(x)$ with respect to the number of iterations in the training process. The value of $\Phi(x)$ can be calculated because $y^*(x)$ has a closed form in this problem and the projection operation onto a simplex is also available to compute. The code is available at https://github.com/WH-XIAN/AS-SGDA.

The experimental results are shown in Figure 1. SGDA-1 and SGDA-2 are the results of SGDA with two largest learning rates from $\{0.1, 0.01, 0.001, 0.0001, 1e-5, 1e-6\}$ that make it converge. G-SGDA is the result of our Generalized SGDA method with the suitable stepsize strategy. From the results in Figure 1 we can see our suitable stepsize strategy improves the convergence speed or stability of SGDA algorithm significantly on all datasets, which validates the effectiveness of our Generalized SGDA method.

6. Conclusion

In this paper we investigate the convergence analysis of minimax optimization algorithms under the relaxation of Lipschitz smooth condition. We provide some counterexamples to reveal that non-Lipschitz smoothness and divergence issues could occur in minimax problems. We propose some generalized minimax algorithms with the suitable stepsize strategy to tackle this issue. We prove that variants of fundamental minimax optimization algorithms GDA, SGDA, GDmax and SGDmax can still converge under the generalized Lipschitz smooth conditions and achieve the same gradient complexity or SFO complexity as their counterparts do in the Lipschitz smooth case. We conduct a numerical experiment of robust logistic regression task to validate the practical performance of our methods.

Acknowledgement

This work was partially supported by NSF IIS 2347592, 2347604, 2348159, 2348169, DBI 2405416, CCF 2348306, CNS 2347617.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. Robust optimization for non-convex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Chen, Z., Zhou, Y., Liang, Y., and Lu, Z. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. *arXiv preprint arXiv:2303.02854*, 2023.

Drenth, J. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass auc optimization. In *International conference on machine learning*, pp. 906–914. PMLR, 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Hao, J., Gong, X., and Liu, M. Bilevel optimization under unbounded smoothness: A new algorithm and convergence analysis, 2024.

Huang, F., Wu, X., and Huang, H. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34: 10431–10443, 2021.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *The Journal of Machine Learning Research*, 23(1):1616–1685, 2022.

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pp. 4880–4889. PMLR, 2020.

Jin, J., Zhang, B., Wang, H., and Wang, L. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34:1792–1804, 2021.

Li, H., Qian, J., Tian, Y., Rakhlin, A., and Jadbabaie, A. Convex and non-convex optimization under generalized smoothness. *arXiv preprint arXiv:2306.01264*, 2023.

Li, J., Zhu, L., and So, A. M.-C. Nonsmooth composite nonconvex-concave minimax optimization. *arXiv preprint arXiv:2209.10825*, 2022.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020a.

Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pp. 2738–2779. PMLR, 2020b.

Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. Min-max optimization without gradients: Convergence and applications to adversarial ml. *arXiv preprint arXiv:1909.13806*, 2019.

Luo, L., Ye, H., and Zhang, T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Neural Information Processing Systems (NeurIPS)*, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Miao, J., Charalambous, P., Kirz, J., and Sayre, D. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

Yan, Y., Xu, Y., Lin, Q., Zhang, L., and Yang, T. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv:1904.10112*, 2019.

Yang, J., Zhang, S., Kiyavash, N., and He, N. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pp. 482–492. PMLR, 2021.

A. Convergence Analysis of Generalized GDA

First, we will provide the proof for the following essential Lemmas.

Lemma A.1. *For any x and x' that satisfy $\|x'-x\| \leq \frac{G}{l_x(\|\nabla_x f(x,y^*(x))\|+G)}$ for some $G \geq 0$, we have*

$$\|y^*(x) - y^*(x')\| \leq \kappa \|x - x'\| \tag{19}$$

Proof. Since $y^*(\cdot)$ is the maximizer, for $\forall y \in \mathcal{Y}$ we have

$$\langle y - y^*(x), \nabla_y f(x, y^*(x))\rangle \leq 0,\ \langle y - y^*(x'), \nabla_y f(x', y^*(x'))\rangle \leq 0 \tag{20}$$

Sum these two inequalities together and we can obtain

$$\langle y^*(x) - y^*(x'), \nabla_y f(x', y^*(x')) - \nabla_y f(x, y^*(x))\rangle \leq 0 \tag{21}$$

As function f is strongly concave with respect to y, we have

$$\mu\|y^*(x) - y^*(x')\|^2 \leq \langle y^*(x) - y^*(x'), \nabla_y f(x', y^*(x')) - \nabla_y f(x', y^*(x))\rangle \tag{22}$$

Combine above two inequalities and we achieve

$$\mu\|y^*(x) - y^*(x')\|^2 \leq \langle y^*(x) - y^*(x'), \nabla_y f(x, y^*(x)) - \nabla_y f(x', y^*(x))\rangle \tag{23}$$

When $\|x'-x\| \leq \frac{G}{l_x(\|\nabla_x f(x,y^*(x))\|+G)}$ for some $G > 0$, by Assumption 4.3 we have

$$\|\nabla_y f(x, y^*(x)) - \nabla_y f(x', y^*(x))\| \leq l_y(\|\nabla_y f(x, y^*(x))\|) \cdot \|x - x'\| \tag{24}$$

When $\mathcal{Y} = \mathbb{R}^{d_2}$, we have $\|\nabla_y f(x, y^*(x))\| = 0$. As function $l_y(\cdot)$ is non-decreasing, we have $l_y(0) \leq l_y(4G_y)$. When f is Lipschitz smooth with respect to y, function $l_y(\cdot)$ is constant L_y and we still have $l_y(2G_y) = L_y$. Combine Eq. (23) and (24), we can reach the conclusion in Lemma A.1. □

Lemma A.2. *For any x and x' that satisfy $\|x'-x\| \leq \frac{G}{l_x(\|\nabla\Phi(x)\|+G)}$ for some $G \geq 0$, we have*

$$\|\nabla\Phi(x) - \nabla\Phi(x')\| \leq 2\kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x - x'\| \tag{25}$$

$$\Phi(x') \leq \Phi(x) + \langle \nabla\Phi(x), x' - x\rangle + \kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x - x'\|^2 \tag{26}$$

$$\Phi(x') \geq \Phi(x) + \langle \nabla\Phi(x), x' - x\rangle - \kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x - x'\|^2 \tag{27}$$

Proof. By Lemma A.1 and Assumption 4.3 we have

$$\begin{aligned}\|\nabla\Phi(x') - \nabla\Phi(x)\| &= \|\nabla_x f(x', y^*(x')) - \nabla_x f(x, y^*(x))\| \\ &\leq l_x(\|\nabla\Phi(x)\| + G) \cdot (\|x' - x\| + \|y^*(x') - y^*(x)\|) \\ &\leq 2\kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x' - x\|\end{aligned} \tag{28}$$

Hence for any $z(t) = x + t(x' - x)$, we have

$$\|\nabla\Phi(z(t)) - \nabla\Phi(x)\| \leq 2t\kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x' - x\| \tag{29}$$

Since we have the equation

$$\Phi(x') = \Phi(x) + \langle \nabla\Phi(x), x' - x\rangle + \int_0^1 \langle \nabla\Phi(z(t)) - \Phi(x), x' - x\rangle dt \tag{30}$$

we can obtain

$$\begin{aligned}\|\Phi(x') - \Phi(x) - \langle \nabla\Phi(x), x' - x\rangle\| &\leq 2\kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x' - x\|^2 \cdot \int_0^1 t dt \\ &\leq \kappa l_x(\|\nabla\Phi(x)\| + G) \cdot \|x - x'\|^2\end{aligned} \tag{31}$$

which leads to the last two inequalities in Lemma A.2. □

Lemma A.3. *For any x, we have*

$$\|\nabla\Phi(x)\|^2 \le 4\kappa l_x(2\|\nabla\Phi(x)\|)\cdot(\Phi(x)-\Phi^*) \tag{32}$$

Proof. Let $x' = x - \frac{\nabla\Phi(x)}{2\kappa l_x(2\|\nabla\Phi(x)\|)}$. By Lemma A.2 we have

$$\Phi^* \le \Phi(x') \le \Phi(x) - \frac{\|\nabla\Phi(x)\|^2}{4\kappa l_x(2\|\nabla\Phi(x)\|)} \tag{33}$$

which implies the conclusion of Lemma A.3. □

Lemma A.4. *For $\forall x$ and y, we have*

$$\|\nabla_y f(x,y)\|^2 \le 2\cdot l_y(2\|\nabla_y f(x,y)\|)\cdot(f(x,y^*(x)) - f(x,y)) \tag{34}$$

Proof. By Assumption 4.3 and the definition of maximizer $y^*(\cdot)$ we have

$$\begin{aligned} f(x,y^*(x)) &\ge f(x, y + \frac{\nabla_y f(x,y)}{l_y(2\|\nabla_y f(x,y)\|)}) \\ &\ge f(x,y) + \frac{1}{l_y(2\|\nabla_y f(x,y)\|)}\|\nabla_y f(x,y)\|^2 - \frac{1}{2\cdot l_y(2\|\nabla_y f(x,y)\|)}\|\nabla_y f(x,y)\|^2 \\ &= f(x,y) + \frac{1}{2\cdot l_y(2\|\nabla_y f(x,y)\|)}\|\nabla_y f(x,y)\|^2 \end{aligned} \tag{35}$$

which implies the conclusion in Lemma A.4. □

Lemma A.5. *Let $\eta_t = \frac{\eta}{S_t} \le \frac{C_0}{16\kappa^2 l_x(2G_x)}$, $\eta_y = \frac{1}{l_y(2G_y)}$ and $\|y_0 - y_0^*\| \le \frac{C_0 G_x}{l_x(2G_x)}$. When $\mathcal{Y} = \mathbb{R}^{d_2}$, we have $\|\nabla\Phi(x_t)\| \le G_x$, $\|\nabla_x f(x_t,y_t)\| \le 2G_x$, $\|\nabla_y f(x_t,y_t)\| \le G_y$ and $\|y_t - y_t^*\| \le \frac{C_0 G_x}{l_x(2G_x)}$ for all $t \ge 0$, where constant $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$.*

Proof. We apply mathematical induction to prove the conclusions in Lemma A.5. According to Lemma A.3 and the definition of G_x, we have $\|\nabla\Phi(x_0)\| \le G_x$. As $\|y_0 - y_0^*\| \le \frac{G_x}{l_x(2G_x)} \le \frac{G_x}{l_x(\|\nabla\Phi(x_0)\|+G_x)}$, by Assumption 4.3 we can further obtain

$$\|\nabla_x f(x_0,y_0) - \nabla\Phi(x_0)\| \le l_x(2G_x)\cdot\|y_0 - y_0^*\| \le G_x \tag{36}$$

which implies $\|\nabla_x f(x_0,y_0)\| \le 2G_x$. Hence the conditions of case $t=0$ are satisfied.

Assume that the conclusions are satisfied for case $t \le \tau$. When $t = \tau+1$, by the requirement of η_t we have

$$\eta_\tau\|\nabla_x f(x_\tau,y_\tau)\| \le 2\eta_\tau G_x \le \frac{G_x}{l_x(2G_x)} \le \frac{G_x}{l_x(\|\nabla\Phi(x_\tau)\|+G_x)} \tag{37}$$

where we have used the induction assumption $\|\nabla_x f(x_\tau,y_\tau)\| \le 2G_x$ and $\|\nabla\Phi(x_\tau)\| \le G_x$. Then we can apply Lemma A.1 and Lemma A.2 to achieve

$$\|y_{\tau+1}^* - y_\tau^*\| \le \kappa\|x_{\tau+1} - x_\tau\| \tag{38}$$

and

$$\begin{aligned} \Phi(x_{\tau+1}) &\le \Phi(x_\tau) + \langle\nabla\Phi(x_\tau), x_{\tau+1} - x_\tau\rangle + \kappa l_x(2G_x)\cdot\|x_{\tau+1} - x_\tau\|^2 \\ &= \Phi(x_\tau) - \eta_\tau\langle\nabla\Phi(x_\tau), \nabla_x f(x_\tau,y_\tau)\rangle + \kappa l_x(2G_x)\cdot\eta_\tau^2\|\nabla_x f(x_\tau,y_\tau)\|^2 \\ &= \Phi(x_\tau) - \frac{\eta_\tau}{2}\|\nabla\Phi(x_\tau)\|^2 + \frac{\eta_\tau}{2}\|\nabla_x f(x_\tau,y_\tau) - \nabla\Phi(x_\tau)\|^2 - \frac{\eta_\tau}{2}(1 - 2\kappa l_x(2G_x)\cdot\eta_\tau)\|\nabla_x f(x_\tau,y_\tau)\|^2 \\ &\le \Phi(x_\tau) - \frac{\eta_\tau}{2}\|\nabla\Phi(x_\tau)\|^2 + \frac{\eta_\tau}{2}\|\nabla_x f(x_\tau,y_\tau) - \nabla\Phi(x_\tau)\|^2 \end{aligned} \tag{39}$$

where the last inequality is obtained by the condition of η_τ. Next we will prove $\|y_{\tau+1} - y^*_{\tau+1}\| \leq \frac{G_x}{l_x(2G_x)}$. According to the update rule of y and the non-expansion property of projection, we have

$$\begin{aligned}\|y^*_\tau - y_{\tau+1}\|^2 &= \|y^*_\tau - \Pi_{\mathcal{Y}}(y_\tau + \eta_y \nabla_y f(x_\tau, y_\tau))\|^2 \\ &\leq \|y^*_\tau - y_\tau - \eta_y \nabla_y f(x_\tau, y_\tau)\|^2 \\ &= \|y^*_\tau - y_\tau\|^2 - 2\eta_y \langle \nabla_y f(x_\tau, y_\tau), y^*_\tau - y_\tau \rangle + \eta_y^2 \|\nabla_y f(x_\tau, y_\tau)\|^2 \end{aligned} \tag{40}$$

As function f is strongly-concave with respect to y, we have

$$\langle \nabla_y f(x_\tau, y_\tau), y^*_\tau - y_\tau \rangle \geq \frac{\mu}{2}\|y^*_\tau - y_\tau\|^2 + f(x_\tau, y^*_\tau) - f(x_\tau, y_\tau) \tag{41}$$

Combine Eq. (40), (41) and Lemma A.4, we have

$$\begin{aligned}\|y^*_\tau - y_{\tau+1}\|^2 &\leq (1 - \mu\eta_y)\|y^*_\tau - y_\tau\|^2 - 2\eta_y(1 - \eta_y \cdot l_y(2G_y))(f(x_\tau, y^*_\tau) - f(x_\tau, y_\tau)) \\ &\leq (1 - \mu\eta_y)\|y^*_\tau - y_\tau\|^2 \leq (1 - \frac{1}{\kappa})\|y^*_\tau - y_\tau\|^2 \end{aligned} \tag{42}$$

where we have used the induction assumption $\|\nabla_y f(x_\tau, y_\tau)\| \leq G_y$ and $\eta_y = \frac{1}{l_y(2G_y)}$. Combine Eq. (38), (42), the induction assumption $\|y_\tau - y^*_\tau\| \leq \frac{G_x}{l_x(2G_x)}$ and $\|\nabla_x f(x_\tau, y_\tau)\| \leq 2G_x$, we have

$$\begin{aligned}\|y^*_{\tau+1} - y_{\tau+1}\| &\leq \|y^*_\tau - y_{\tau+1}\| + \|y^*_{\tau+1} - y^*_\tau\| \\ &\leq (1 - \frac{1}{2\kappa})\|y^*_\tau - y_\tau\| + \kappa\eta_\tau\|\nabla_x f(x_\tau, y_\tau)\| \leq (1 - \frac{1}{2\kappa})\frac{G_x}{l_x(2G_x)} + 2\kappa\eta_\tau G_x \leq \frac{G_x}{l_x(2G_x)} \end{aligned} \tag{43}$$

where we have used the requirement of η_τ in the last inequality. As $\|y_\tau - y^*_\tau\| \leq \frac{G_x}{l_x(2G_x)} \leq \frac{G_x}{l_x(\|\nabla\Phi(x_\tau)\| + G_x)}$, by Assumption 4.3 we can obtain

$$\|\nabla_x f(x_\tau, y_\tau) - \nabla\Phi(x_\tau)\|^2 \leq l_x^2(2G_x) \cdot \|y_\tau - y^*_\tau\|^2 \tag{44}$$

By Young's inequality we have

$$\begin{aligned}\|y_\tau - y^*_\tau\|^2 &\leq (1 + \frac{1}{2\kappa - 1})\|y^*_{\tau-1} - y_\tau\|^2 + 2\kappa\|y^*_\tau - y^*_{\tau-1}\|^2 \\ &\leq \frac{2\kappa}{2\kappa - 1} \cdot \frac{\kappa - 1}{\kappa}\|y_{\tau-1} - y^*_{\tau-1}\|^2 + 2\kappa^3\eta^2_{\tau-1}\|\nabla_x f(x_{\tau-1}, y_{\tau-1})\|^2 \\ &\leq (1 - \frac{1}{2\kappa} + 4\kappa^3\eta^2_{\tau-1}l_x^2(2G_x))\|y_{\tau-1} - y^*_{\tau-1}\|^2 + 4\kappa^3\eta^2_{\tau-1}\|\nabla\Phi(x_{\tau-1})\|^2 \\ &\leq (1 - \frac{1}{4\kappa})\|y_{\tau-1} - y^*_{\tau-1}\|^2 + 4\kappa^3\eta^2_{\tau-1}\|\nabla\Phi(x_{\tau-1})\|^2 \end{aligned} \tag{45}$$

where the second inequality is derived by the same way as Eq. (42) and (38); the third inequality is derived by Cauchy-Schwartz inequality and Assumption 4.3; the last inequality is derived by the condition of η_t. Let $\gamma = 1 - \frac{1}{4\kappa}$. Applying recursion to Eq. (45), we can obtain

$$\|y_\tau - y^*_\tau\|^2 \leq \gamma^\tau\|y_0 - y^*_0\|^2 + 4\kappa^3\sum_{s=0}^{\tau-1}\gamma^{\tau-1-s}\eta_s^2\|\nabla\Phi(x_s)\|^2 \tag{46}$$

Inserting Eq. (44) and (46) into (39), we have

$$\Phi(x_{\tau+1}) \leq \Phi(x_\tau) - \frac{\eta_\tau}{2}\|\nabla\Phi(x_\tau)\|^2 + \frac{\eta_\tau l_x^2(2G_x)}{2}\Big(\gamma^\tau\|y_0 - y^*_0\|^2 + 4\kappa^3\sum_{s=0}^{\tau-1}\gamma^{\tau-1-s}\eta_s^2\|\nabla\Phi(x_s)\|^2\Big) \tag{47}$$

Applying recursion to above inequality, we can achieve

$$\begin{aligned}\Phi(x_{\tau+1}) &\leq \Phi(x_0) - \sum_{t=0}^{\tau} \frac{\eta_t}{2}\Big(1 - 4\kappa^3\eta_t^2 l_x^2(2G_x)\sum_{s=t}^{\tau-1}\gamma^{s-t}\Big)\|\nabla\Phi(x_t)\|^2 + \frac{l_x^2(2G_x)\cdot\|y_0 - y_0^*\|^2}{2}\sum_{t=0}^{\tau}\gamma^t\eta_t \\ &\leq \Phi(x_0) - \sum_{t=0}^{\tau}\frac{\eta_t}{2}\Big(1 - 16\kappa^4\eta_t^2 l_x^2(2G_x)\Big)\|\nabla\Phi(x_t)\|^2 + \frac{G_x^2}{32\kappa^2 l_x(2G_x)}\sum_{t=0}^{\tau}\gamma^t \\ &\leq \Phi(x_0) - \sum_{t=0}^{\tau}\frac{15\eta_t}{32}\|\nabla\Phi(x_t)\|^2 + \frac{G_x^2}{8\kappa l_x(2G_x)} \end{aligned} \tag{48}$$

where we have used the setup of η_t and $\|y_0 - y_0^*\|$. According to the definition of G_x, we have

$$\Phi(x_{\tau+1}) - \Phi^* \leq (\Phi(x_0) - \Phi^*) - \sum_{t=0}^{\tau}\frac{15\eta_t}{32}\|\nabla\Phi(x_t)\|^2 + (\Phi(x_0) - \Phi^*) \leq 2(\Phi(x_0) - \Phi^*) \tag{49}$$

Combining Eq. (49), Lemma A.3 and the definition of G_x, we can reach the conclusion that $\|\nabla\Phi(x_{\tau+1})\| \leq G_x$. As $\|y_{\tau+1} - y_{\tau+1}^*\| \leq \frac{G_x}{l_x(2G_x)} \leq \frac{G_x}{l_x(\|\nabla\Phi(x_{\tau+1})\|+G_x)}$, by Assumption 4.3 we can obtain

$$\|\nabla_x f(x_{\tau+1}, y_{\tau+1}) - \nabla\Phi(x_{\tau+1})\| \leq l_x(2G_x)\cdot\|y_{\tau+1} - y_{\tau+1}^*\| \leq G_x \tag{50}$$

which implies $\|\nabla_x f(x_{\tau+1}, y_{\tau+1})\| \leq 2G_x$. Finally, we need to estimate $\|\nabla_y f(x_{\tau+1}, y_{\tau+1})\|$. We have

$$\begin{aligned}\|\nabla_y f(x_{\tau+1}, y_{\tau+1})\| &= \|\nabla_y f(x_{\tau+1}, y_{\tau+1}) - \nabla_y f(x_{\tau+1}, y_{\tau+1}^*)\| \\ &\leq l_y(0)\cdot\|y_{\tau+1} - y_{\tau+1}^*\| \leq \frac{C_0 l_y(0) G_x}{l_x(2G_x)} \leq G_y \end{aligned} \tag{51}$$

which is obtained by the definition of constant C_0. □

Lemma A.6. *Let $\eta_t = \frac{\eta}{S_t} \leq \frac{C_0}{16\kappa^2 l_x(2G_x)}$, $\eta_y = \frac{1}{l_y(2G_y)}$ and $\|y_0 - y_0^*\| \leq \frac{C_0 G_x}{l_x(2G_x)}$. When $l_y(\cdot) \equiv L_y$, we have $\|\nabla\Phi(x_t)\| \leq G_x$, $\|\nabla_x f(x_t, y_t)\| \leq 2G_x$ and $\|y_t - y_t^*\| \leq \frac{C_0 G_x}{l_x(2G_x)}$ for all $t \geq 0$, where constant $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(0)G_x}\}$.*

Proof. Different from the case $\mathcal{Y} = \mathbb{R}^{d_2}$, we do not need the upper bound for $\nabla_y f(x_t, y_t)$. In Lemma A.5 the only place that needs the condition is Eq. (42), which requires $l_y(2\|\nabla_y f(x_\tau, y_\tau)) \leq l_y(2G_y)$. However, when f is Lipschitz smooth with respect to y, this condition is always satisfied since $l_y(\cdot) \equiv L_y$. The rest part of proof is the same as Lemma A.5. □

With Lemma A.5, Lemma A.6 and Eq. (49), we can reach the conclusion in Theorem 4.4. When $S_t \equiv S$, the result of Corollary 4.5 can be directly achieved by Theorem 4.4. Next, we will prove other corollaries using different options to compute S_t. By Cauchy-Schwartz inequality, we have the following conclusion based on Theorem 4.4.

Lemma A.7. *Suppose the conditions in Theorem 4.4 are satisfied. Then we have*

$$(\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|)^2 \leq \frac{5(\Phi(x_0) - \Phi^*)(\sum_{t=0}^{T-1} S_t)}{\eta} \tag{52}$$

We also need the following Lemma A.8

Lemma A.8. *Suppose the conditions in Theorem 4.4 are satisfied. Then we have*

$$\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\| \leq 2\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\| + 4\kappa G_x \tag{53}$$

Proof. By the proof of Lemma A.5 we have

$$\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\| \leq l_x(2G_x)\cdot\|y_t - y_t^*\| \tag{54}$$

By Eq. (42) we have

$$\begin{aligned}\|y_t - y_t^*\| &\le \|y_{t-1}^* - y_t\| + \|y_t^* - y_{t-1}^*\| \le (1 - \frac{1}{2\kappa}) \cdot \|y_{t-1} - y_{t-1}^*\| + \kappa\eta_{t-1}\|\nabla_x f(x_{t-1}, y_{t-1})\| \\ &\le (1 - \frac{1}{2\kappa} + \kappa\eta_{t-1} l_x(2G_x)) \cdot \|y_{t-1} - y_{t-1}^*\| + \kappa\eta_{t-1}\|\nabla\Phi(x_{t-1})\| \\ &\le (1 - \frac{1}{4\kappa}) \cdot \|y_{t-1} - y_{t-1}^*\| + \kappa\eta_{t-1}\|\nabla\Phi(x_{t-1})\| \end{aligned} \tag{55}$$

Let $\gamma = 1 - \frac{1}{4\kappa}$. Applying recursion to above inequality and we can obtain

$$\|y_t - y_t^*\| \le \gamma^t \|y_0 - y_0^*\| + \kappa \sum_{s=0}^{t-1} \gamma^{t-1-s} \eta_s \|\nabla\Phi(x_s)\| \tag{56}$$

Summing Eq. (54) and combining with Eq. (56), we achieve

$$\begin{aligned}\sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\| &\le 4\kappa l_x(2G_x) \cdot \|y_0 - y_0^*\| + \kappa l_x(2G_x) \sum_{t=0}^{T-1} \eta_t \|\nabla\Phi(x_t)\| \cdot \sum_{s=t}^{T-1} \gamma^{s-t} \\ &\le 4\kappa G_x + \frac{1}{4} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\| \end{aligned} \tag{57}$$

Hence we can reach the conclusion of Lemma A.8. □

When S_t is computed by option (1) or (4), we have $\sum_{t=0}^{T-1} S_t \le \sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\| + T\epsilon$ and

$$(\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\|)^2 \le \frac{10(\Phi(x_0) - \Phi^*)}{\eta T} (\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\|) + \frac{\epsilon(\Phi(x_0) - \Phi^*)}{T} + \frac{4\kappa G_x(\Phi(x_0) - \Phi^*)}{T^2} \tag{58}$$

The first term on the right side is the dominant term when we have $\eta = O(\frac{\epsilon}{\kappa^2})$ and $T = O(\kappa^2\epsilon^{-2})$. We have

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla\Phi(x_t)\| \le \frac{20(\Phi(x_0) - \Phi^*)}{\eta T} \tag{59}$$

which implies the result in Corollary 4.6. When S_t is computed by option (2), we have we have $\sum_{t=0}^{T-1} S_t \le \log(T) \sum_{t=0}^{T-1} \|\nabla_x f(x_t, y_t)\| + T\epsilon$. Mimic above steps and we can achieve Corollary 4.7. Therefore, we have completed the convergence analysis of the Generalized GDA algorithm.

B. Convergence Analysis of Generalized GDmax

Lemma B.1. *Let* $\eta_t = \frac{\eta}{S_t} \le \frac{C_0}{16\kappa l_x(2G_x)}$, $\eta_y = \frac{1}{l_y(4G_y)}$, $K = \kappa \log(\frac{1}{\theta})$ *and* $\|y_0 - y_0^*\| \le \frac{C_0 G_x}{l_x(2G_x)}$. *When* $\mathcal{Y} = \mathbb{R}^{d_2}$, *we have* $\|\nabla\Phi(x_t)\| \le G_x$, $\|\nabla_x f(x_t, y_t)\| \le 2G_x$, $\|\nabla_y f(x_t, y_t)\| \le G_y$ *and* $\|y_t - y_t^*\| \le \frac{C_0 G_x}{l_x(2G_x)}$ *for all* $t \ge 0$, *where constant* $C_0 = \min\{1, \frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$ *and* $\theta = \min\{\frac{1}{4}, \frac{1}{l_x^2(2G_x)}\}$.

Proof. It is easy to check that the following result in Lemma A.5 are still satisfied.

$$\begin{aligned}\Phi(x_{\tau+1}) &\le \Phi(x_\tau) - \frac{\eta_\tau}{2}\|\nabla\Phi(x_\tau)\|^2 + \frac{\eta_\tau}{2}\|\nabla_x f(x_\tau, y_\tau) - \nabla\Phi(x_\tau)\|^2 \\ &\le \Phi(x_\tau) - \frac{\eta_\tau}{2}\|\nabla\Phi(x_\tau)\|^2 + \frac{\eta_\tau l_x^2(2G_x)}{2}\|y_\tau - y_\tau^*\|^2 \end{aligned} \tag{60}$$

The difference is the way to estimate $\|y_\tau - y_\tau^*\|$. As we have

$$\eta_\tau\|\nabla_x f(x_\tau, y_\tau)\| \le 2\eta_\tau G_x \le \frac{G_x}{l_x(2G_x)} \le \frac{G_x}{l_x(\|\nabla\Phi(x_\tau)\| + G_x)} \tag{61}$$

by Assumption 4.3 we can achieve

$$\begin{aligned}\|\nabla_y f(x_{\tau+1},y_\tau)\| &\le \|\nabla_y f(x_{\tau+1},y_\tau)-\nabla_y f(x_\tau,y_\tau)\|+\|\nabla_y f(x_\tau,y_\tau)\|\\ &\le l_y(2G_y)\cdot\eta_\tau\|\nabla_x f(x_\tau,y_\tau)\|+G_y\le 2G_y\end{aligned}\tag{62}$$

where we have used the definition of η_τ and constant C_0 in the last inequality. $y_{\tau+1}$ is computed via a nested loop, which can be regarded as a strongly-convex minimization subproblem starting at $-f(x_{\tau+1},y_\tau)$. According to the result in minimization problem (theorem 4.3 in (Li et al., 2023)), when we set $\eta_y=\frac{1}{l_y(4G_y)}$ and $K=\kappa\log(\frac{1}{\theta})$, we have

$$\begin{aligned}\|y_{\tau+1}-y^*_{\tau+1}\|^2 &\le \theta\|y_\tau-y^*_{\tau+1}\|^2\le 2\theta\|y^*_{\tau+1}-y^*_\tau\|^2+2\theta\|y_\tau-y^*_\tau\|^2\\ &\le 2\theta\kappa^2\eta_\tau^2\|\nabla_x f(x_\tau,y_\tau)\|^2+2\theta\|y_\tau-y^*_\tau\|^2\le\frac{C_0^2G_x^2}{l_x^2(2G_x)}\end{aligned}\tag{63}$$

where the last inequality is achieved when $\theta\le\frac{1}{4}$. From Eq. (63) we can also obtain

$$\begin{aligned}\|y_{\tau+1}-y^*_{\tau+1}\|^2 &\le (2\theta+4\theta\kappa^2\eta_\tau^2l_x^2(2G_x))\|y_\tau-y^*_\tau\|^2+4\theta\kappa^2\eta_\tau^2l_x^2(2G_x)\|\nabla\Phi(x_\tau)\|^2\\ &\le 3\theta\|y_\tau-y^*_\tau\|^2+\frac{\theta}{64}\|\nabla\Phi(x_\tau)\|^2\end{aligned}\tag{64}$$

where we have used the setup of η_t to simplify the inequality. Applying recursion to Eq. (60) and (64), we can obtain

$$\begin{aligned}\Phi(x_{\tau+1}) &\le \Phi(x_0)-\sum_{t=0}^{\tau}\frac{\eta_t}{2}\Big(1-\frac{\theta l_x^2(2G_x)}{16}\Big)\|\nabla\Phi(x_t)\|^2+\frac{G_x^2}{8\kappa l_x(2G_x)}\\ &\le \Phi(x_0)-\sum_{t=0}^{\tau}\frac{15\eta_t}{32}\|\nabla\Phi(x_t)\|^2+\frac{G_x^2}{8\kappa l_x(2G_x)}\end{aligned}\tag{65}$$

where we have used $\theta\le\min\{\frac{1}{4},\frac{1}{l_x^2(2G_x)}\}$, $\|y_0-y_0^*\|\le\frac{G_x}{l_x(2G_x)}$ and $\eta_t\le\frac{1}{16\kappa l_x(2G_x)}$. According to the definition of G_x and Lemma A.3,, we can obtain $\|\nabla\Phi(x_{\tau+1})\|\le G_x$. As $\|y_{\tau+1}-y^*_{\tau+1}\|\le\frac{G_x}{l_x(2G_x)}\le\frac{G_x}{l_x(\|\nabla\Phi(x_{\tau+1})\|+G_x)}$, by Assumption 4.3 we can obtain

$$\|\nabla_x f(x_{\tau+1},y_{\tau+1})-\nabla\Phi(x_{\tau+1})\|\le l_x(2G_x)\cdot\|y_{\tau+1}-y^*_{\tau+1}\|\le G_x\tag{66}$$

which implies $\|\nabla_x f(x_{\tau+1},y_{\tau+1})\|\le 2G_x$. Finally, we need to estimate $\|\nabla_y f(x_{\tau+1},y_{\tau+1})\|$. We have

$$\begin{aligned}\|\nabla_y f(x_{\tau+1},y_{\tau+1})\| &= \|\nabla_y f(x_{\tau+1},y_{\tau+1})-\nabla_y f(x_{\tau+1},y^*_{\tau+1})\|\\ &\le l_y(0)\cdot\|y_{\tau+1}-y^*_{\tau+1}\|\le\frac{C_0l_y(0)G_x}{l_x(2G_x)}\le G_y\end{aligned}\tag{67}$$

which is obtained by the definition of constant C_0. Hence we have finished the mathematical induction. □

Lemma B.2. *Let $\eta_t=\frac{\eta}{S_t}\le\frac{C_0}{16\kappa l_x(2G_x)}$, $\eta_y=\frac{1}{l_y(4G_y)}$, $K=\kappa\log(\frac{1}{\theta})$ and $\|y_0-y_0^*\|\le\frac{C_0G_x}{l_x(2G_x)}$. When $l_y(\cdot)\equiv L_y$, we have $\|\nabla\Phi(x_t)\|\le G_x$, $\|\nabla_x f(x_t,y_t)\|\le 2G_x$ and $\|y_t-y_t^*\|\le\frac{C_0G_x}{l_x(2G_x)}$ for all $t\ge 0$, where constant $C_0=\min\{1,\frac{l_x(2G_x)G_y}{l_y(2G_y)G_x}\}$ and $\theta=\min\{\frac{1}{4},\frac{1}{l_x^2(2G_x)}\}$.*

Proof. Different from Lemma B.1, in this case we do not need to estimate an upper bound for $\nabla_y f(x_t,y_t)$. In the case of Lemma B.1, the upper bound of $\nabla_y f(x_t,y_t)$ is required because when solving the l_y-smooth strongly-convex subproblem, the stepsize and complexity are affected by the initial gradient norm. But when the function is Lipschitz smooth, the requirement is unnecessary and we can set the stepsize to $\eta_y=\frac{1}{L_y}$. □

Based on Lemma B.1, Lemma B.2 and Eq. (65), we can reach the conclusions of Theorem 4.8 and Corollary 4.9. Mimic the steps of Lemma A.7 and Lemma A.8, we can prove the results in Corollary 4.10 and Corollary 4.11.

C. Convergence Analysis of Generalized SGDA

In stochastic algorithms, we need the following auxiliary Lemmas.

Lemma C.1. *Let vector X be a stochastic variable. Then we have*

$$0 \leq \mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2 \tag{68}$$

Lemma C.2. *Let $X_1, X_2, \cdots, X_n$ be n independent stochastic variables of which the means are 0. Then we have*

$$\mathbb{E}\|\sum_{i=1}^{n} X_i\|^2 = \sum_{i=1}^{n} \mathbb{E}\|X_i\|^2 \tag{69}$$

Next, we will provide the proof for Theorem 4.13. Here we will only consider the case of $\mathcal{Y} = \mathbb{R}^{d_2}$ because the operations for the case $l_y(\cdot) \equiv L_y$ is similar to deterministic algorithms. For convenience, we denote $\eta_t = \frac{\eta}{S_t}$. Recall that in the stochastic case constant G_x is re-defined as follows:

$$G_x = \max\{u > 0 | u^2 \leq 32\kappa l_x(2u) \cdot (\Phi(x_0) - \Phi^* + \sigma^2)/\delta\}$$

Proof. First, we define

$$T_0 = \min\{ \min\{t \mid \Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta} \text{ or } \|y_t - y_t^*\| > \frac{C_0 G_x}{l_x(2G_x)}\}, T\} \tag{70}$$

We will prove that the probability of $T_0 < T$ is small. According to the definition of G_x and T_0, we know when $t < T_0$, we have $\|\nabla\Phi(x_t)\| \leq G_x$ and $\|y_t - y_t^*\| \leq \frac{C_0 G_x}{l_x(2G_x)}$. From the proof of Lemma A.5, it can also be checked that $\|\nabla_x f(x_t, y_t)\| \leq 2G_x$ and $\|\nabla_y f(x_t, y_t)\| \leq G_y$. By the update rule of y we have

$$\begin{aligned}
\|y_t^* - y_{t+1}\|^2 &= \|y_t^* - \Pi_{\mathcal{Y}}(y_t + \eta_y u_t)\|^2 \\
&\leq \|y_t^* - y_t - \eta_y u_t\|^2 \\
&= \|y_t^* - y_t\|^2 - 2\eta_y\langle u_t - \nabla_y f(x_t, y_t), y_t^* - y_t - \eta_y \nabla_y f(x_t, y_t)\rangle - 2\eta_y\langle \nabla_y f(x_t, y_t), y_t^* - y_t\rangle \\
&\quad + \eta_y^2\|\nabla_y f(x_t, y_t)\|^2 + \eta_y^2\|u_t - \nabla_y f(x_t, y_t)\|^2
\end{aligned} \tag{71}$$

When $t < T_0$, taking expectation on ξ_t, by Eq. (41), Lemma A.4 and Lemma C.2 we have

$$\mathbb{E}\|y_t^* - y_{t+1}\|^2 \leq (1 - \frac{1}{\kappa})\mathbb{E}\|y_t^* - y_t\|^2 + \frac{\eta_y^2\sigma^2}{b_y} \tag{72}$$

Hence by Young's inequality we have

$$\begin{aligned}
\mathbb{E}\|y_{t+1}^* - y_{t+1}\|^2 &\leq (1 + \frac{1}{2\kappa - 1})\mathbb{E}\|y_t^* - y_{t+1}\|^2 + 2\kappa\mathbb{E}\|y_{t+1}^* - y_t^*\|^2 \\
&\leq (1 - \frac{1}{2\kappa})\mathbb{E}\|y_t^* - y_t\|^2 + 2\kappa^3\eta_t^2\mathbb{E}\|v_t\|^2 + \frac{\eta_y^2\sigma^2}{b_y} \\
&\leq (1 - \frac{1}{2\kappa} + 6\kappa^3\eta_t^2 l_x^2(2G_x))\mathbb{E}\|y_t^* - y_t\|^2 + 6\kappa^3\eta_t^2\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{6\kappa^3\eta_t^2\sigma^2}{b_x} + \frac{\eta_y^2\sigma^2}{b_y} \\
&\leq (1 - \frac{1}{4\kappa})\mathbb{E}\|y_t^* - y_t\|^2 + 6\kappa^3\eta_t^2 G_x^2 + \frac{6\kappa^3\eta_t^2\sigma^2}{b_x} + \frac{\eta_y^2\sigma^2}{b_y}
\end{aligned} \tag{73}$$

Applying recursion and the setup of η_t, we can achieve

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq (1 - \frac{1}{4\kappa})^t\|y_0^* - y_0\|^2 + \frac{\delta C_0^2 G_x^2}{96 l_x^2(2G_x)} + \frac{\delta C_0^2\sigma^2}{96 l_x^2(2G_x) b_x} + \frac{4\kappa\eta_y^2\sigma^2}{b_y} \leq \frac{\delta C_0^2 G_x^2}{16 l_x^2(2G_x)} \tag{74}$$

for $t \leq T_0$ where we have used $b_x \geq \frac{\sigma^2}{G_x^2}$, $b_y \geq \frac{192\kappa\sigma^2 l_x^2(2G_x)}{\delta G_x^2 l_y^2(2G_y)}$ and the condition of $\|y_0^* - y_0\|$.

Mimic the steps in Eq. (39), we can also obtain

$$\begin{aligned}\Phi(x_{t+1}) &\leq \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t\rangle + \kappa l_x(2G_x)\cdot\|x_{t+1} - x_t\|^2 \\ &= \Phi(x_t) - \eta_t\langle \nabla\Phi(x_t), v_t\rangle + \kappa l_x(2G_x)\cdot\eta_t^2\|v_t\|^2 \end{aligned} \tag{75}$$

for $t < T_0$. Taking expectation on ξ_t, by Lemma C.2 we have

$$\begin{aligned}\mathbb{E}\Phi(x_{t+1}) &\leq \mathbb{E}\Phi(x_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\eta_t}{2}\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla\Phi(x_t)\|^2 + \kappa l_x(2G_x)\cdot\eta_t^2\mathbb{E}\|v_t - \nabla_x f(x_t,y_t)\|^2 \\ &\quad - \frac{\eta_t}{2}(1 - 2\kappa l_x(2G_x)\cdot\eta_t)\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 \\ &\leq \mathbb{E}\Phi(x_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\eta_t}{2}\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla\Phi(x_t)\|^2 + \frac{\kappa l_x(2G_x)\cdot\eta_t^2\sigma^2}{b_x} \\ &\leq \mathbb{E}\Phi(x_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\eta_t l_x^2(2G_x)}{2}\mathbb{E}\|y_t^* - y_t\|^2 + \frac{\kappa l_x(2G_x)\cdot\eta_t^2\sigma^2}{b_x} \end{aligned} \tag{76}$$

From Eq. (73) we can also achieve

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq (1 - \frac{1}{4\kappa})\mathbb{E}\|y_{t-1}^* - y_{t-1}\|^2 + 6\kappa^3\eta_t^2\mathbb{E}\|\nabla\Phi(x_{t-1})\|^2 + \frac{6\kappa^3\eta_t^2\sigma^2}{b_x} + \frac{\eta_y^2\sigma^2}{b_y} \tag{77}$$

Let $\gamma = 1 - \frac{1}{\kappa}$ and apply recursion to Eq. (77), then we can obtain

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq \gamma^t\|y_0^* - y_0\|^2 + 6\kappa^3\sum_{s=0}^{t-1}\gamma^{t-1-s}\eta_s^2\mathbb{E}\|\nabla\Phi(x_s)\|^2 + \frac{\delta^2 C_0^2\sigma^2}{96 l_x^2(2G_x)b_x} + \frac{4\kappa\eta_y^2\sigma^2}{b_y} \tag{78}$$

Insert Eq. (78) into Eq. (76) and summing over t. We have

$$\begin{aligned}\mathbb{E}\Phi(x_{t+1}) &\leq \Phi(x_0) - \sum_{s=0}^{t-1}\frac{\eta_s}{2}\Big(1 - \frac{\delta^2 C_0^2}{96}\Big)\mathbb{E}\|\nabla\Phi(x_s)\|^2 + \frac{\delta C_0 l_x(2G_x)}{24\kappa}\|y_0^* - y_0\|^2 + \frac{\delta^2 C_0^2\sigma^2(t+1)}{2304\kappa^3 l_x(2G_x)b_x} \\ &\quad + \frac{\delta^3 C_0^3\sigma^2(t+1)}{9216\kappa^2 l_x(2G_x)b_x} + \frac{\eta_y^2 l_x(2G_x)\sigma^2(t+1)}{24\kappa b_y} \end{aligned} \tag{79}$$

As $T = \frac{\kappa^2}{\delta^2\epsilon^2}$, $b_x \geq \frac{\sigma^2}{G_x^2\epsilon^2}$, $b_y \geq \max\{\frac{192\kappa\sigma^2 l_x^2(2G_x)}{\delta G_x^2 l_y^2(2G_y)}, \frac{\kappa l_x(2G_x)}{\delta^2 l_y^2(2G_y)\epsilon^2}\}$, we have

$$\mathbb{E}\Phi(x_{t+1}) \leq \Phi(x_0) - \sum_{s=0}^{t-1}\frac{95\eta_s}{192}\mathbb{E}\|\nabla\Phi(x_s)\|^2 + \frac{\delta G_x^2}{32\kappa l_x(2G_x)} + \sigma^2 \tag{80}$$

According to the definition of G_x, for all $t \leq T_0$ we have

$$\mathbb{E}\Phi(x_t) - \Phi^* \leq 2(\Phi(x_0) - \Phi^* + \sigma^2) \tag{81}$$

If $T_0 < T$, then we have $\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta}$ or $\|y_t - y_t^*\| > \frac{C_0 G_x}{l_x(2G_x)}$ at $t = T_0$. According to Markov's inequality and Eq. (74), we have

$$Pr(\|y_t - y_t^*\|^2 > \frac{C_0^2 G_x^2}{l_x^2(2G_x)}|t = T_0) \leq \mathbb{E}\|y_t^* - y_t\|^2/(\frac{C_0^2 G_x^2}{l_x^2(2G_x)}) \leq \frac{\delta}{4} \tag{82}$$

According to Markov's inequality and Eq. (81), we have

$$Pr(\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta}|t = T_0) \leq (\mathbb{E}\Phi(x_t) - \Phi^*)/\frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta} \leq \frac{\delta}{4} \tag{83}$$

By union bound we have

$$Pr(T_0 < T) \leq Pr(\|y_t - y_t^*\|^2 > \frac{C_0^2 G_x^2}{l_x^2(2G_x)}|t = T_0) + Pr(\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta}|t = T_0) \leq \frac{\delta}{2} \tag{84}$$

If $T_0 = T$, by Eq. (80) we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\mathbb{E}\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{5(\Phi(x_0)-\Phi^*+\sigma^2)}{\eta T} \tag{85}$$

By Markov's inequality, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{10(\Phi(x_0)-\Phi^*+\sigma^2)}{\delta\eta T} \tag{86}$$

with probability at least $1-\frac{\delta}{2}$. By union bound, we can finish the proof of Theorem 4.13. □

When $S_t \equiv S$, we can set $\eta_t = \frac{\delta C_0}{48\kappa^2 l_x(2G_x)}$. By Theorem 4.13, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\Phi(x_t)\|^2 \leq \frac{480(\Phi(x_0)-\Phi^*+\sigma^2)\epsilon^2}{C_0} \tag{87}$$

which reaches the conclusion of Corollary 4.14. Mimic the steps in Lemma A.7 and Lemma A.8, we can prove the results in Corollary 4.15 and Corollary 4.16.

D. Convergence Analysis of Generalized SGDmax

In this section we will provide the proof for Theorem 4.17. Here we will only consider the case of $\mathcal{Y} = \mathbb{R}^{d_2}$ because the operations for the case $l_y(\cdot) \equiv L_y$ is similar to deterministic algorithms. For convenience, we denote $\eta_t = \frac{\eta}{S_t}$.

Proof. Similar to the analysis of SGDA, we define

$$T_0 = \min\{\ \min\{t \mid \Phi(x_t)-\Phi^* > \frac{8(\Phi(x_0)-\Phi^*+\sigma^2)}{\delta} \text{ or } \|y_t - y_t^*\| > \frac{C_0 G_x}{l_x(2G_x)}\}, T\} \tag{88}$$

We will prove that the probability of $T_0 < T$ is small. When $t < T_0$, according the proof of Lemma B.1 it can be checked that all induction assumptions still hold. Hence we still have

$$\mathbb{E}\Phi(x_{t+1}) \leq \mathbb{E}\Phi(x_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\eta_t l_x^2(2G_x)}{2}\mathbb{E}\|y_t^* - y_t\|^2 + \frac{\kappa l_x(2G_x)\cdot\eta_t^2\sigma^2}{b_x} \tag{89}$$

as what we have done in Eq. (76). According to the update rule of y, we have

$$\begin{aligned}
&\|y_t^* - y_{t-1,k+1}\|^2 \\
&= \|y_t^* - \Pi_{\mathcal{Y}}(y_{t-1,k} + \eta_y u_{t-1,k})\|^2 \\
&\leq \|y_t^* - y_{t-1,k} - \eta_y u_{t-1,k}\|^2 \\
&= \|y_t^* - y_{t-1,k}\|^2 - 2\eta_y\langle u_{t-1,k} - \nabla_y f(x_t, y_{t-1,k}), y_t^* - y_{t-1,k} - \eta_y\nabla_y f(x_t, y_{t-1,k})\rangle \\
&\quad - 2\eta_y\langle\nabla_y f(x_t, y_{t-1,k}), y_t^* - y_{t-1,k}\rangle + \eta_y^2\|\nabla_y f(x_t, y_{t-1,k})\|^2 + \eta_y^2\|u_{t-1,k} - \nabla_y f(x_t, y_{t-1,k})\|^2
\end{aligned} \tag{90}$$

Taking expectation and we achieve

$$\mathbb{E}\|y_t^* - y_{t-1,k+1}\|^2 \leq (1-\frac{1}{\kappa})\mathbb{E}\|y_t^* - y_{t-1,k}\|^2 + \frac{\eta_y^2\sigma^2}{b_y} \tag{91}$$

Apply recursion to above inequality and we can obtain

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq (1-\frac{1}{\kappa})^K\mathbb{E}\|y_t^* - y_{t-1}\|^2 + \frac{\kappa\eta_y^2\sigma^2}{b_y} \tag{92}$$

As $K = \kappa \log(\frac{1}{\theta})$, we have

$$\begin{aligned}\mathbb{E}\|y_t^* - y_t\|^2 &\leq \theta \mathbb{E}\|y_t^* - y_{t-1}\|^2 + \frac{\kappa \eta_y^2 \sigma^2}{b_y} \\ &\leq 2\theta \mathbb{E}\|y_{t-1}^* - y_{t-1}\|^2 + 2\theta \mathbb{E}\|y_t^* - y_{t-1}^*\|^2 + \frac{\kappa \eta_y^2 \sigma^2}{b_y} \\ &\leq 2\theta \mathbb{E}\|y_{t-1}^* - y_{t-1}\|^2 + 2\theta \kappa^2 \eta_{t-1}^2 \mathbb{E}\|v_{t-1}\|^2 + \frac{\kappa \eta_y^2 \sigma^2}{b_y} \\ &\leq (2\theta + 6\theta \kappa^2 \eta_{t-1}^2 l_x^2(2G_x)) \mathbb{E}\|y_{t-1}^* - y_{t-1}\|^2 + 6\theta \kappa^2 \eta_{t-1}^2 \mathbb{E}\|\nabla \Phi(x_{t-1})\|^2 + \frac{6\theta \kappa^2 \eta_{t-1}^2 \sigma^2}{b_x} + \frac{\kappa \eta_y^2 \sigma^2}{b_y} \end{aligned} \tag{93}$$

Since we have $\eta_t \leq \frac{\delta C_0}{48 \kappa l_x(2G_x)}$, $b_x \geq \frac{\sigma^2}{G_x^2}$, $b_y \geq \frac{24 \kappa \sigma^2 l_x(2G_x)}{\delta G_x^2 l_y^2(4G_y)}$, $\theta \leq \frac{1}{4}$ and $\|\nabla \Phi(x_t)\| \leq G_x$ when $t < T_0$, we can obtain

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq \|y_0^* - y_0\|^2 + \frac{\delta^2 C_0^2 G_x^2}{48 l_x^2(2G_x)} + \frac{\delta^2 C_0^2 G_x^2}{48 l_x^2(2G_x)} + \frac{\delta C_0^2 G_x^2}{6 l_x^2(2G_x)} \leq \frac{\delta C_0^2 G_x^2}{4 l_x^2(2G_x)} \tag{94}$$

for $t \leq T_0$. Besides, from Eq. (93) we can also obtain

$$\mathbb{E}\|y_t^* - y_t\|^2 \leq (\frac{3}{4})^t \|y_0^* - y_0\|^2 + \frac{\theta \delta^2 C_0^2}{384 l_x^2(2G_x)} \sum_{s=0}^{t-1} (\frac{3}{4})^{t-1-s} \mathbb{E}\|\nabla \Phi(x_s)\|^2 + \frac{\theta \delta^2 C_0^2 \sigma^2}{96 l_x^2(2G_x) b_x} + \frac{4 \kappa \eta_y^2 \sigma^2}{b_y} \tag{95}$$

Combining with Eq. (89) and summing over t, we achieve

$$\begin{aligned}\mathbb{E}\Phi(x_t) \leq \Phi(x_0) - \sum_{s=0}^{t-1} \frac{\eta_s}{2} (1 - \frac{\delta^2 C_0^2}{384}) \mathbb{E}\|\nabla \Phi(x_s)\|^2 + \frac{\delta C_0 l_x(2G_x)}{24\kappa} \|y_0^* - y_0\|^2 + \frac{\delta^2 C_0^2 \sigma^2 t}{2304 \kappa l_x(2G_x) b_x} \\ + \frac{\delta^3 C_0^3 \sigma^2 t}{9216 \kappa l_x(2G_x) b_x} + \frac{\eta_y^2 l_x(2G_x) \sigma^2 t}{24 b_y}\end{aligned} \tag{96}$$

for $t \leq T_0$. As $T = \frac{\kappa}{\delta^2 \epsilon^2}$, $b_x \geq \frac{\sigma^2}{G_x^2 \epsilon^2}$, $b_y \geq \max\{\frac{24 \kappa \sigma^2 l_x^2(2G_x)}{\delta G_x^2 l_y^2(4G_y)}, \frac{\kappa l_x(2G_x)}{\delta^2 l_y^2(4G_y) \epsilon^2}\}$, we have

$$\mathbb{E}\Phi(x_{t+1}) \leq \Phi(x_0) - \sum_{s=0}^{t-1} \frac{95 \eta_s}{192} \mathbb{E}\|\nabla \Phi(x_s)\|^2 + \frac{\delta G_x^2}{32 \kappa l_x(2G_x)} + \sigma^2 \tag{97}$$

According to the definition of G_x, for all $t \leq T_0$ we have

$$\mathbb{E}\Phi(x_t) - \Phi^* \leq 2(\Phi(x_0) - \Phi^* + \sigma^2) \tag{98}$$

If $T_0 < T$, then we have $\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta}$ or $\|y_t - y_t^*\| > \frac{C_0 G_x}{l_x(2G_x)}$ at $t = T_0$. According to Markov's inequality and Eq. (94), we have

$$Pr(\|y_t - y_t^*\|^2 > \frac{C_0^2 G_x^2}{l_x^2(2G_x)} | t = T_0) \leq \mathbb{E}\|y_t^* - y_t\|^2 / (\frac{C_0^2 G_x^2}{l_x^2(2G_x)}) \leq \frac{\delta}{4} \tag{99}$$

According to Markov's inequality and Eq. (98), we have

$$Pr(\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta} | t = T_0) \leq (\mathbb{E}\Phi(x_t) - \Phi^*) / \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta} \leq \frac{\delta}{4} \tag{100}$$

By union bound we have

$$Pr(T_0 < T) \leq Pr(\|y_t - y_t^*\|^2 > \frac{C_0^2 G_x^2}{l_x^2(2G_x)} | t = T_0) + Pr(\Phi(x_t) - \Phi^* > \frac{8(\Phi(x_0) - \Phi^* + \sigma^2)}{\delta} | t = T_0) \leq \frac{\delta}{2} \tag{101}$$

If $T_0 = T$, by Eq. (97) we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\mathbb{E}\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{5(\Phi(x_0)-\Phi^*+\sigma^2)}{\eta T} \tag{102}$$

By Markov's inequality, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\|\nabla\Phi(x_t)\|^2}{S_t} \leq \frac{10(\Phi(x_0)-\Phi^*+\sigma^2)}{\delta\eta T} \tag{103}$$

with probability at least $1-\frac{\delta}{2}$. By union bound, we can finish the proof of Theorem 4.17. □

The rest proof for Corollary 4.18, Corollary 4.19 and Corollary 4.20 is similar to the analysis of SGDA. Hence we will omit that part of proof to avoid redundancy.